Benutzerorientierte Bewertungsmaßstäbe für Information Retrieval Systeme: Der *Robust Task* bei CLEF 2006

Thomas Mandl

Universität Hildesheim
Informationswissenschaft
Marienburger Platz 22
31141 Hildesheim
mandl@uni-hildesheim.de

Zusammenfassung

Die Qualität von Antworten im Information Retrieval schwankt zwischen einzelnen Anfragen sehr stark. Die Evaluierung im Information Retrieval zielt in der Regel auf eine Optimierung der durschnittlichen Retrieval-Qualität über mehrere Testanfragen (Topics). Sehr schlecht beantwortete Anfragen wirken sich besonders negativ auf die Zufriedenheit des Benutzers aus. Neue Ansätze zur Evaluierung der Robustheit von Systemen werten daher die schwierigen Anfragen stärker. Im Rahmen des Cross Language Evaluation Forum (CLEF) wurde 2006 ein *Robust Task* durchgeführt. Der Artikel zeigt die Gründe für Entwicklung dieser Aufgabenstellung nach, referiert die Ergebnisse und verweist auf zukünftige Planungen.

Abstract:

The quality of the responses of information retrieval system differs greatly between queries. The evaluation in information retrieval usually considers the average retrieval quality over several queries. However, queries which lead to poor results have a higher impact on the satisfaction of the user. New appraoches try to consider the users perspective by emphazising the hard topics. Within the Cross Language Evaluation Forum (CLEF), a robust task has been established in 2006. This article reports the reasons for the development, the task design and shows results and potenntial future trends.

1 Einleitung

Die Evaluierung im Information Retrieval zielt in der Regel auf eine Optimierung der durschnittlichen Retrieval-Qualität über mehrere Testanfragen (Topics). Der Eindruck des Benutzers von einem System wird aber stark durch sehr schlecht beantwortete Anfragen geprägt. Wird eine Nullantwort durch eine geringfügige Verbesserung für ein Topic vermieden, so hilft dies dem Benutzer meist mehr als eine geringfügige Steigerung bei einer ohnehin gut beantworteten Anfrage. Die *Mean Average Precision* (MAP) als wichtigstes Maß bei allen Evaluierungen weist beiden Fällen die gleiche Bedeutung zu.

Neuere Entwicklungen in der Evaluierungsforschung versuchen, diese Benutzerperspektive in das Zentrum zu rücken. Vor diesem Hintergrund ist die Debatte um adäquate Evaluierungsmaße im Information Retrieval wieder aufgegriffen worden. Zwar ist bekannt, dass die meisten Evaluierungsmaße eine starke Korrelation untereinander aufweisen (BUCKLEY & VOORHEES 2005). Trotzdem ist die Analyse benutzerorientierter Maßzahlen sinnvoll.

Ein wichtiges Maß, welches häufig eingesetzt wird und das die schwierigeren Anfragen stärker gewichtet, ist der geometrische Durchschnitt oder der geometrische Mittelwert. Er berechnet sich als die n-te Wurzel aus dem Produkt der zu mittelnden Einzelwerte.

$$geoAve = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

Die Einzelwerte stellen im Information Retrieval die Ergebnisse der einzelnen Aufgaben (Topics, Anfragen) dar. Denkbar wäre auch, die Ergebnisse der einzelnen Topics bereits anders zu gewichten. Viele Benutzer, vor allem bei Internet-Suchmaschinen bewerten die Precision unter den ersten Treffern besonders hoch und legen geringeren Wert auf den Recall.

Für die Bewertung der Robustheit hat sich besonders im Rahmen der Text Retrieval Conference (TREC) das geometrische Mittel etabliert. Abbildung 1 zeigt ein konstruiertes Beispiel, welches dessen Wirkung verdeutlicht.

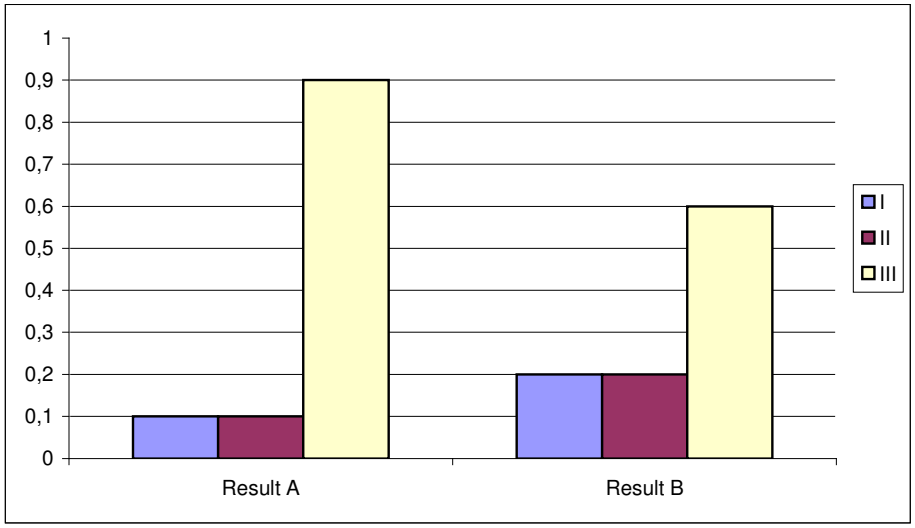

Abb. 1: Exemplarische Ergebnisse zweier Retrieval Systeme

Tabelle 1: Auswertung der Ergebnisse aus Abbildung 1

Topic	System	Ergebnis	Topic	System	Ergebnis
1	A	0,1	1	B	0,2
2	A	0,1	2	B	0,2
3	A	0,9	3	B	0,6
GeoAve	A	0,21	GeoAve	B	**0,29**
MAP	A	**0,37**	MAP	B	0,33

Abbildung 1 und Tabelle 1 gegen die Ergebnisse von zwei Systemen für drei Anfragen (Topics) wider. System A erreicht bei Topic 3 ein optimales Ergebnis und bei Topic 1 und 2 nur ein schlechtes Ergebnis. Während das System A bei der Bewertung durch die durchschnittliche Precision besser abschneidet, weist das geometrische Mittel System B die bessere Gesamtbewertung zu. Da der Benutzer in den meisten Situationen die Verbesserung von System B für Topic 1 und 2 höher bewerten wird, als die Verschlechterung auf hohem Niveau bei Topic 3, entspricht das geometrische Mittel eher der Benutzer-Perspektive.

Es zeigt sich in vielen Evaluierungen, dass Ergebnisse wie das oben konstruierte durchaus vorkommen (siehe Abschnitt 3).

2 Robustheit als Ziel im Information Retrieval

Die Variabilität zwischen den Topics ist bei allen Evaluierungen meist größer als die zwischen den Systemen (BUCKELY 2004, BRASCHLER 2003). Dementsprechend verspricht die Analyse der einzelnen Topics großes Potential für die Verbesserung der Retrieval-Ergebnisse (MANDL & WOMSER-HACKER 2003, HARMAN & BUCKLEY 2004):

> „More work needs to be done on customizing methods for each topic“ (HARMAN 2005)

Für die Steigerung der Robustheit ist es erforderlich, besonders die Qualität der Systeme für die schwierigen Topics zu erhöhen. Dazu ist sowohl die automatische Vorhersage der Schwierigkeit als auch die Analyse und besondere Behandlung der vermutlich schwierigen Topics nötig.

Der Workshop *Reliable Information Access* (RIA)[1] versuchte das zweite Problem zu bearbeiten (HARMAN 2004). RIA wurde vom das *National Institute for Standards and Technology* (NIST) in Gaithersburg, Maryland, USA veranstaltet und dort durchgeführt. Führende Forschungsgruppen trafen am NIST zusammen und arbeiteten dort mit ihren Systemen vor Ort an den besonders schwierigen Topics aus der TREC Evaluierungsinitiative.

Die Text Retrieval Conference (TREC[2]) findet seit 1992 statt und hat eine neue Ära in der Evaluierung im Information Retrieval eingeläutet. Im Jahr 2004 veranstaltete das *National Institute for Standards and Technology* (NIST) bereits den dreizehnten Workshop dieser ersten großen Evaluierungsinitiative (VOORHEES & BUCKLAND 2004). Der ad-hoc Track war zu Beginn von TREC der wichtigste und bedeutendste Track (WOMSER-HACKER 2004) und fand bis 1999 statt. Danach wurde ad-hoc Retrieval in anderen Tracks forgeführt. Eine Variante ist der *Robust Retrieval Track*, der 2003, 2004 und 2005 durchgeführt wurde. Die Datengrundlage für den Robust Track bei TREC 2003 bestand aus ca. 500.000 Dokumenten aus alten TREC ad-hoc Tracks. Die Bewertungsmethodologie wurde modifiziert und die Gruppen wurden informiert, dass die Bewertung hauptsächlich anhand des geometri-

[1] http://ir.nist.gov/ria/
[2] http://trec.nist.gov

schen Mittelwerts erfolgen würde. Dadurch richteten die Organisatoren des *Robust Retrieval Track* die Evaluierungsmethodik stärker an den Bedürfnissen des Benutzers aus. Der Fokus lag nicht auf der durchschnittlichen Performanz der Systeme über alle Topics, sondern auf einer stabilen Performanz über alle Topics. Dazu mussten Ergebnisse für schwierige Topics stärker werten als einzelne herausragende Ergebnisse für eher leichte Topics, was durch die geometrische Mittelwertbildung erreicht wurde (VOORHEES 2005a, VOORHEES 2005b). Eine weitere Aufgabe im *Robust Retrieval Track* bestand darin, die Topics nach ihrer Schwierigkeit zu ordnen.

Der RIA Workshop widmete sich der Analyse der Gründe für das Scheitern der Systeme bei bestimmten Topics. Aus dem RIA Workshop entstand der *Robust Track* im Rahmen von TREC.

Ein weiterer kürzlich durchgeführter Workshop widmete sich dem Thema der Schwierigkeit einzelner Topics. Der Workshop *Predicting Query Difficulty - Methods and Applications* bei SIGIR 2005[3] versuchte, den Schwierigkeitsgrad einzeler Anfragen vorherzusagen.

3 Robust Task bei CLEF 2006

Basierend auf den Erfahrungen bei TREC wurde auch im Rahmen von CLEF ein *Robust Task* durchgeführt, der vom Autor organisiert wurde.

Die CLEF-Initiative[4] (MANDL 2005, PETERS ET AL. 2005) etablierte sich im Jahr 2000. Seitdem steigt die Zahl der Teilnehmer bei CLEF stetig an und die Forschung zu mehrsprachigen Information Retrieval Systemen gewinnt an Dynamik. CLEF folgt dem Modell von TREC und schuf eine mehrsprachige Kollektion mit Zeitungstexten. Inzwischen umfasst die Dokument-Kollektion für das ad-hoc Retrieval die Sprachen Englisch, Französisch, Spanisch, Italienisch, Deutsch, Holländisch, Schwedisch, Finnisch, Portugiesisch, Bulgarisch, Ungarisch und Russisch. Mehrere weitere Tracks wie *Question Answering*, *Web-Retrieval*, *Spoken Dokument Retrieval* oder *Geographic CLEF* untersuchen bestimmte Aspekte des mehrsprachigen Retrieval.

Vor der Einführung eines *Robust Tasks* mussten hierfür die Ergebnisse früherer CLEF Experimente analysiert werden. Die Einführung neuer Evaluierungsmaße lohnt nur, wenn die Korrelation zwischen den traditionellen und den neuer Maßen nicht zu hoch ist. Dazu wurden die Ergebnisse von CLEF 2001 bis 2003 untersucht. Um die Ähnlichkeit der Rankings auf der Basis des MAP und des geometrischen Mittelwerts zu bestimmen, wurde der Spearman Rang-Korrelations-Koeffizient berechnet. Für eine identische Rangfolge liefert der Koeefizient den Wert 1, für eine umgekehrte Rangfolge den Wert –1. Die folgende Tabelle 3 zeigt die Korrelationen für einige ausgewählte Experiment-Typen.

[3] http://www.haifa.ibm.com/sigir05-qp/
[4] http://www.clef-campaign.org

Tabelle 2: Rang Korrelation für CLEF Experimente nach Spearman

Task	**Topic Sprache**	**CLEF Jahr**	**Korrelation**
Mono	German	2001	0,91
Multi	English	2001	0,96
Mono	Spanish	2001	0,93
Bi	English	2002	0,98

Die Korrelationen sind sehr hoch, jedoch sind die Rangfolgen nicht identisch. Bei genauerer Analyse erweisen sich auch relevante Unterschiede. Die Top-Systeme unterscheiden sich in mehreren Fällen. Um dies zu illustrieren, zeigt Tabelle 3 die Position des besten Systems in der MAP Rangfolge auf einer nach dem geometrischen Mittel sortierten Skala.

Tabelle 3: Rang der besten Systeme nach MAP im geoAve Ranking

Task	**Topic Sprache**	**CLEF Jahr**	**Rang**
Mono	German	2001	2
Multi	English	2001	1
Mono	Spanish	2001	1
Bi	English	2002	10

Auch in anderen Fällen ändern sich Positionen dramatisch. Zum Beispiel fällt das zweitbeste System für CLEF 2001 (Topic Sprache Englisch) auf Platz 32 von 56 Teilnehmern.

Bisher wurde als Definition für die Schwierigkeit eines Topics ein niedriger maximaler MAP-Wert genutzt. Dies entspricht dem Vorgehen der meisten Forscher (EGUCHI ET AL. 2002, KWOK 2005, CRONEN-TOWNSEND ET AL. 2002, MOTHE & TANGUY 2005). Die genaue Analyse der CLEF-Ergebnisse weckt aber Zweifel an dieser Perspektive und der Definition. Zum einen kann sowohl das geometrische Mittel aus den Systemen benutzt werden. Dann wird der Einfluss besonders schlechter Systeme erhöht. So könnte eine Menge von Topics identifiziert werden, die sich für viele System als besonders schwierig erweisen. Daneben könnte auch das Ergebnis des besten Systems für dieses Topic genutzt werden. Dies kann als Wert für das Maximum gelten, welches für dieses Topic erzielt werden kann. Die folgende Tabelle 4 zeigt die Größe der Schnittmenge zwischen den beiden alternativen Definitionen und den zehn schwierigsten Topics nach dem durchschnittlichen MAP Wert aller Systeme.

Tabelle 4: Anzahl unterschiedlicher schwieriger Topics

Task	Mono	Multi	Mono	Bi
Topic Sprache	Deutsch	Englisch	Spanisch	Englisch
CLEF Jahr	2001	2001	2001	2002
Geometr. Mittel	2	2	2	2
Bestes System	3	3	2	2

Es erweist sich, dass die Schnittmenge zwischen den Definitionen sehr klein ist. Weitere Definitionen scheinen denkkbar. Die Menge der relevanten Dokumente in der Kollektion mag ein für den Benutzer sehr einleuchtendes Maß darstellen. Interessant sollte neben der absoluten Definition vor allem das Verbesserungspotential sein, wofür die Varianz zwischen den Systemen berücksichtigen werden müsste. Die Frage, welche Topics eigentlich schwer sind, bleibt letztlich offen.

Das Task Design für den *Robust Task* versuchte vor allem einen große Menge von Dokumenten und Topics zu finden, für die ohne weitere Relevanz-Urteile eine Analyse durchgeführt werden konnte. Die Wahl fiel auf die CLEF Jahhre 2001, 2002 und 2003, in denen für eine Reihe von Kernsprachen eine weitegehend konstante Dokumentenkollektion benutzt wurde. Somit standen aus diesen drei Jahren 160 Topics zur Verfügung.

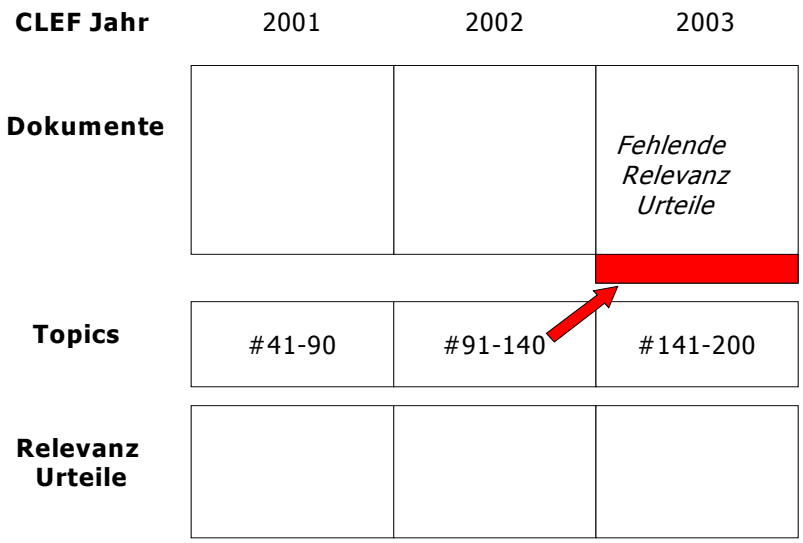

Abbildung 2: Dokumente und Topics für den *Robust Task*

Die folgende Tabelle 5 listet die verwendeten Kollektionen auf. Insgesamt umfasst die sechssprachige Datensammlung 1,35 Millionen Dokumente mit 3,6 Gigabyte Text.

Tabelle 5: Korpora für den *Robust Task*

Sprache	**Kollektion**
Englisch	LA Times 94, Glasgow Herald 95
Französisch	ATS (SDA) 94/95, Le Monde 94
Italienisch	La Stampa 94, AGZ (SDA) 94/95
Holländisch	NRC Handelsblad 94/95, Algemeen Dagblad 94/95
Deutsch	Frankfurter Rundschau 94/95, Spiegel 94/95, SDA 94
Spanisch	EFE 94/95

Die in Abbildung 2 angedeutete Inkonsistenz zwischen Relevanz-Bewertungen und Kollektion entstand durch geringfügige Änderungen der deutschen Kollektion. Die Teilnehmer

sollten das Wissen darüber, dass für einige Topics in der neueren Kollektion keine bewerteten Dokumente enthalten sind, nicht ausnutzen. Dies konnte zu Schwierigkeiten bei der Optimierung der Systeme führen und wurde von einigen Teilnehmern kritisiert. Das Eingrenzen der Ergebnisse für diese Anfragen auf die bewerteten Korpora hätte zu einer Verbesserung von bis zu 10% bei der MAP führen können (SAVOY 2006).

Die Identifikation von besonders schwierigen Topics hatte sich als problematisch erwiesen. Die Schwierigkeit war zwischen den Sprachen sehr unterschiedlich und die Analse bestätigte letztlich ein Ergebnis das Robust Tracks bei TREC. Eine Anfrage ist nicht *per se* schwierig, sondern nur in Zusammenspiel mit einer Kollektion (VOORHEES 2005b).

Somit wurde für den Robust Task bei CLEF keine Menge von schwierigen Topics definiert, sondern die Topics wurden zufällig in zwei Mengen geteilt. Während 60 Topics zum Training dienten, sollten die übrigen 100 als Testdaten benutzt werden.

4 Ergebnisse

Insgesamt betiligten sich acht Gruppen an dem Robust Task und reichten 133 Runs (Experimente) ein (DI NUNZIO ET AL. 2006).

Table 6 : Teilnehmer am CLEF Robust Task 2006 (DI NUNZIO ET AL. 2006)

U. Coruna & U. Sunderland (Spanien & UK)	Hummingbird Core Tech. (Kanada)
U. Jaen (Spanien)	U. Neuchatel (Schweiz)
DAEDALUS & Madrid Univs. (Spanien)	Dublin City U. – Computing (Irland)
U. Salamanca – REINA (Spanien)	U. Hildesheim – Inf. Sci. (Deutschland)

Tabelle 7: Anzahl der eingereichten Ergebnisse für den CLEF *Robust Task* 2006

Task	**Sprache**	**Anzahl Test Runs**	**Anzahl Training Runs**	**Anzahl Gruppen**
mono	en	13	7	6
	fr	18	10	7
	nl	7	3	3
	de	7	3	3
	es	11	5	5
	it	11	5	5
bi	it->es	8	2	3
	fr->nl	4	0	1
	en->de	5	1	2
multi	multi	10	3	4

Die folgenden Tabellen zeigen die Ergebnisse der besten Systeme für mono-linguale Experimente (aus DI NUNZIO ET AL. 2006)

Tabelle 8: Ergebnisse für mono-linguale Experimente beim *Robust Task* CLEF 2006

Track	Teilnehmer Rang					
	1	2	3	4	5	Differenz
Deutsch	hummingbird	colesir	daedalus			1. vs 3.
MAP	48,30%	37,21%	34,06%			41,81%
GMAP	22,53%	14,80%	10,61%			112,35%
Run	humDE06Rtde	CoLesIRdeTst	deFSdeR2S			
Italienisch	hummingbird	reina	dcu	daedalus	colesir	1. vs 5.
MAP	41,94%	38,45%	37,73%	35,11%	32,23%	30,13%
GMAP	11,47%	10,55%	9,19%	10,50%	8,23%	39,37%
Run	humIT06Rtde	reinaITtdtest	dcudescit1005	itFSitR2S	CoLesIRitTst	
Spanisch	hummingbird	reina	dcu	daedalus	colesir	1. vs 5.
MAP	45,66%	44,01%	42,14%	40,40%	40,17%	13,67%
GMAP	23,61%	22,65%	21,32%	19,64%	18,84%	25,32%
Run	humES06Rtde	reinaEStdtest	dcudescsp12075	esFSesR2S	CoLesIResTst	

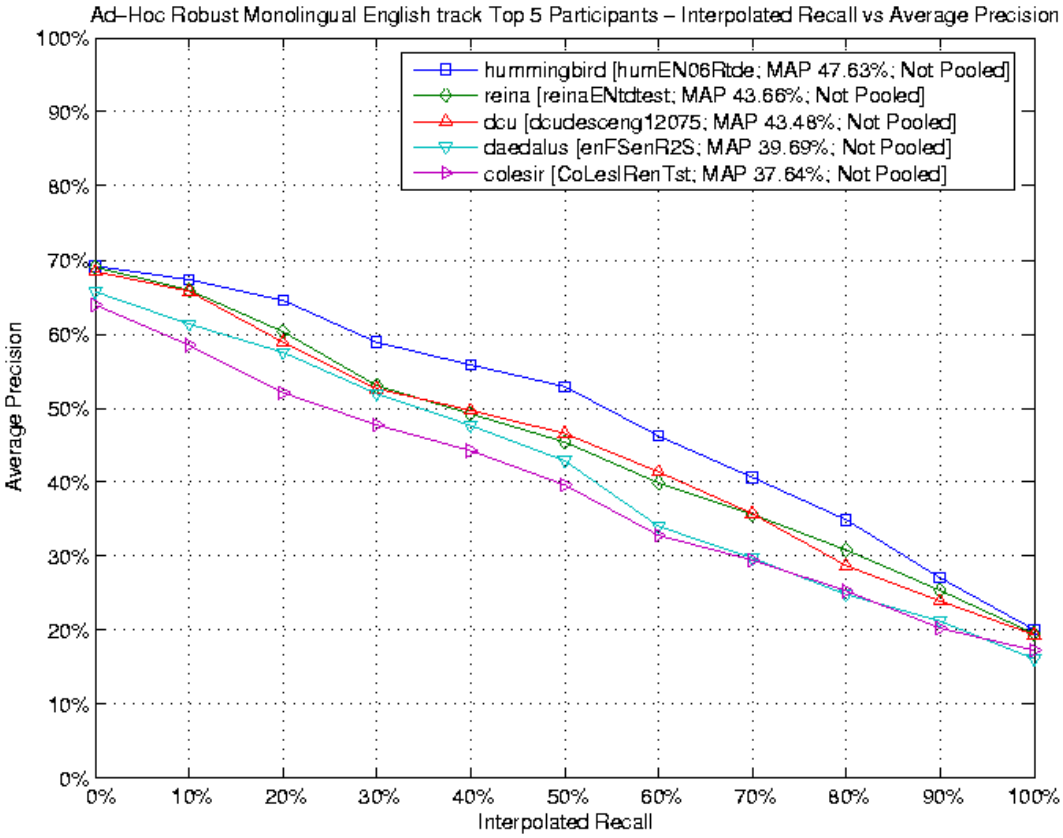

Abb. 3: Ergebnisse des Robust Task bei CLEF 2006 für Englisch (aus DI NUNZIO ET AL. 2006)

Tabelle 9: Ergebnisse für mono-linguale Experimente beim *Robust Task* CLEF 2006

Track	Teilnehmer Rang					
	1	2	3	4	5	Differenz
Holländisch	hummingbird	daedalus	colesir			1. vs 3.
MAP	51,06%	42,39%	41,60%			22,74%
GMAP	25,76%	17,57%	16,40%			57,13%
Run	humNL06Rtde	nlFSnlR2S	CoLesIRnlTst			
Englisch	hummingbird	reina	dcu	daedalus	colesir	1. vs 5.
MAP	47,63%	43,66%	43,48%	39,69%	37,64%	26,54%
GMAP	11,69%	10,53%	10,11%	8,93%	8,41%	39,00%
Run	humEN06Rtde	reinaENtdtest	dcudesceng 12075	enFSenR2S	CoLesIRenTst	
Französisch	unine	hummingbird	reina	dcu	colesir	1. vs 5.
MAP	47,57%	45,43%	44,58%	41,08%	39,51%	20,40%
GMAP	15,02%	14,90%	14,32%	12,00%	11,91%	26,11%
Run	UniNEfrr1	humFR06Rtde	reinaFRtdtest	dcudescfr12075	CoLesIRfrTst	

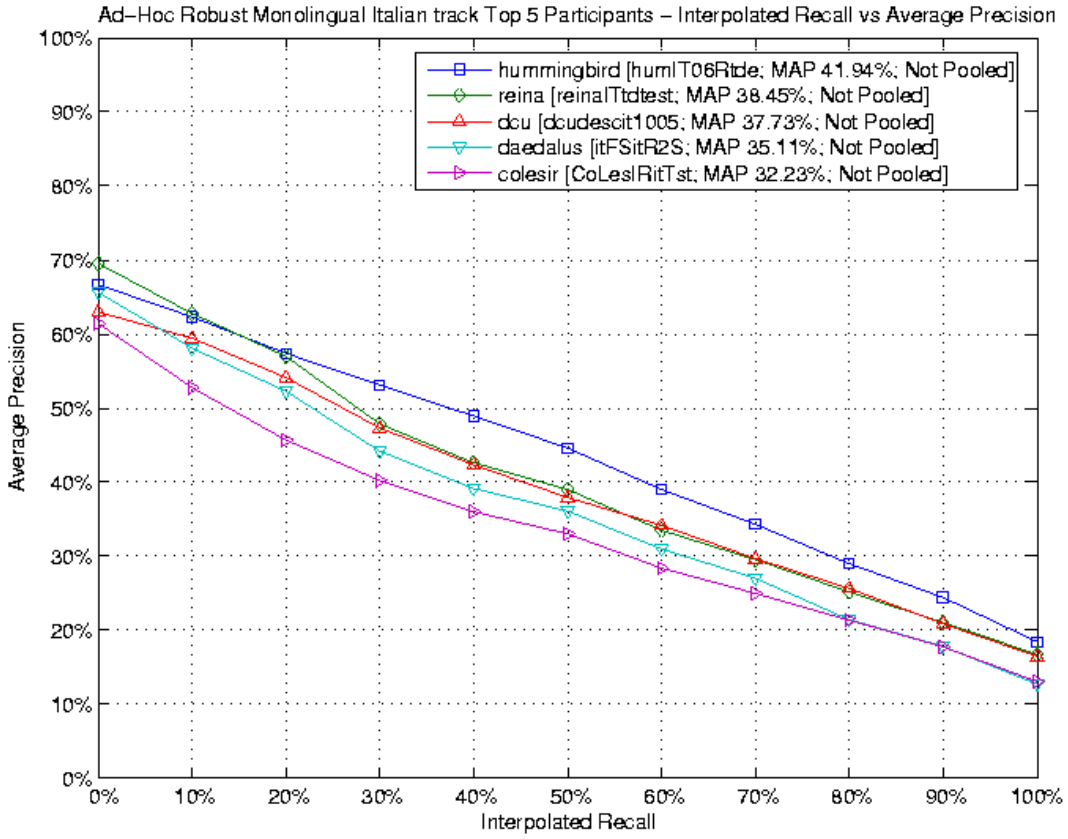

Abb. 4: Ergebnisse des Robust Task bei CLEF 2006 für Italienisch (aus DI NUNZIO ET AL. 2006)

Die Tabellen zeigen, dass die Systeme Systeme sehr gute und vergleichbare Ergebnisse erzielt haben. Die grafische Darstellung der Recall-Precision Kurven unterstreichen dies. Tabelle 10 zeigt die Resultate für die multi-lingualen Experimente der vier teilnehmenden Gruppen.

Tabelle 10: Ergebnisse für multi-linguale Experimente beim *Robust Task* CLEF 2006

Track	Teilnehmer Rang			
	1	2	3	4
Multilingual	jaen	daedalus	colesir	reina
MAP	27,85%	22,67%	22,63%	19,96%
GMAP	15,69%	11,04%	11,24%	13,25%
Run	ujamlrsv2	mlRSFSen2S	CoLesIRmultTst	reinaES2mtdtest

Die Ergebnisse zeigen, dass die Ähnlichkeit zwischen den Rangfolgen nach MAP und GeoAve sehr hoch sind. Lediglich an zwei Stellen ergeben sich Unterschiede, die jedoch nie die oberste Position betreffen. Bei den mehrsprachigen Ergebnissen rückt das vierte System auf Platz 2 vor und bei den mono-lingualen Ergebnissen für Italienisch tauscht das vierte System den Platz mit dem dritten.

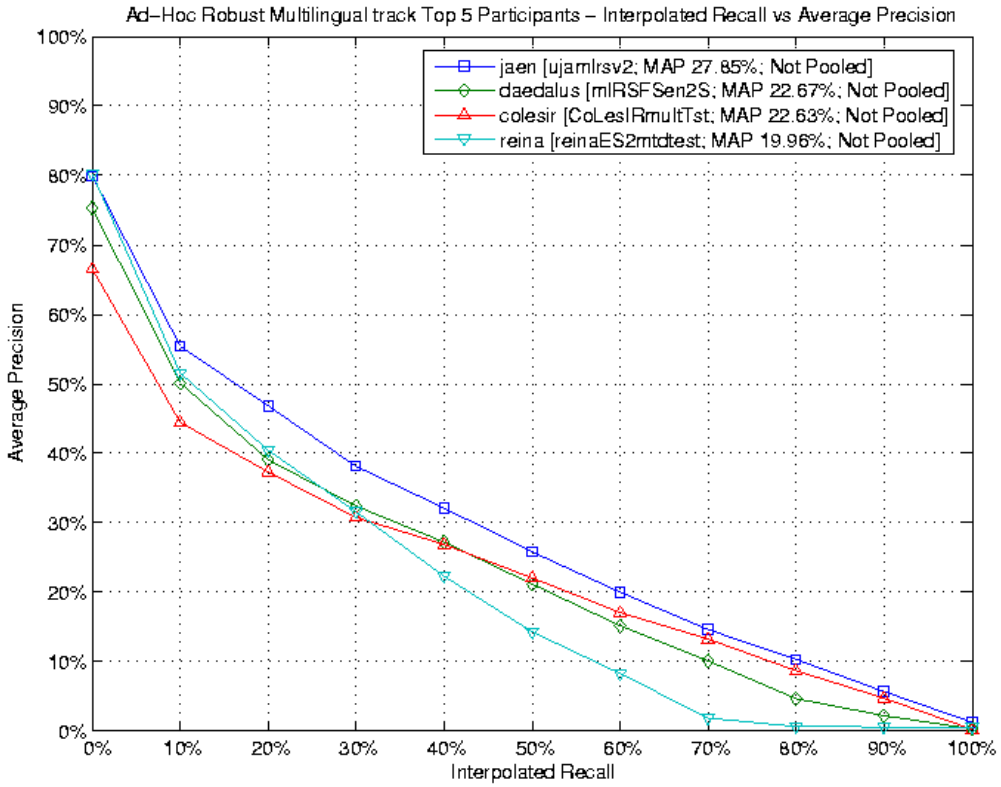

Abb. 5: Ergebnisse des Robust Task für multilinguale Experimente (aus DI NUNZIO ET AL. 2006)

Einige der Teilnehmer verließen sich auf die hohe Korrelation zwischen MAP und geoAVe und optimierten ihre Systeme nicht spezifisch für die robusten Bewertungsmßstäbe. Einige der Forschungsgruppe jedoch griffen Ansätze aus den *Robust Track* von TREC auf. Das SINAI System experimentierte mit der Expansion der Anfrage-Terme mit großen externen Korpora, die bei TREC zu den besten Ergebnissen geführt hatte (VOORHEES 2005a). SINAI nutzte dafür eine Web-Suchmaschine (MARTINEZ-SANTIAGO ET AL. 2006). Das REINA System der Universität Salamanca setzte eine Heuristik für die Identifikation der schwierigen Topics ein. Anschließend wurden für die unterschiedlich schwierigen Anfragen verschiedene Strategien zur Expansion der Anfrage-Terme eingesetzt (ZAZO ET AL. 2006). Hummingbird bewertete seine Experimente mit unterschiedlichen Evaluierungsmaßen. Besonders die Precision nach zehn Dokumenten wurde als besseres Maß als das geometrische Mittel angesehen, das die Benutzerperspektive gut widerspiegelt (TOMLINSON 2006). Das MIRACLE System wendete eine Fusion mehrerer einzelner Systeme an. Die Paramter für die Zusammenführung der individuellen Ergebnisse wurde für die Robustheit optimiert (GONI-MENOYO ET AL. 2006).

Die Analyse der Schwierigkeit der Topics wurde für die Ergebnisse fortgesetzt. Erneut zeigte sich, dass Topics nicht inhärent schwierig sind, sondern nur für bestimmte Kollektionen. Damit erweisen sich für die unterschiedlichen Zielkorpora in den einzelnen Sprachen jeweils andere Topics als sehr schwer. Zur Illustration seien hier einige Beispiele angeführt. Das Topic 64 ist das leichteste für mono-linguales und bi-linguales Retrieval mit Deutsch als Zielsprache. Für Italienisch dagegen ist es das schwerste Topic. Topic 144 führt zu den besten Ergebnissen für bi-linguales Retrieval für holländische Dokumente. Für die Zielsprache Deutsch führt es dagegen zu den schlechtesten Ergebnissen.

5 Planungen für 2007

Die neu vorgeschlagenen Maße führten zu intensiven Diskussionen beim CLEF Workshop. Die Precision nach zehn Dokumenten (TOMLINSON 2006) wurde von vielen Teilnehmern als mögliche Variante begrüßt. Auch die Anzahl der Fehlschläge eines Systems wurde als Alternative erwähnt. Dafür könnte die Anzahl der Topics unter einem gewissen Schwellenwert für die Precision gewertet werden. Hier muss allerding festgestellt werden, dass das ursprüngliche Benutzermodell von TREC von einem Benutzer ausgeht, das sehr viele Dokumente bewertet. Mit der Einführung neuer Maße wird implizit auch ein neues Benutzermodell eingeführt. Dies kann natürlich durchaus sinnvoll sein.

In CLEF 2007 sollen diese Maße eingesetzt werden. Die Datengrundlage soll verändert werden. Für Englisch und Französisch stehen Topics aus sechs Jahren zur Verfügung. Davon sollen drei Jahre für das Training und der Rest für den Test eingesetzt werden. Zusätzlich sollen Topics aus drei Jahren für Portuguieisch angewandt werden, ohne dass hier Trainingsdaten bereit stehen. Als multi-lingualer Task soll bi-linguales Retrieval von Englisch zum Französischen möglich sein.

6 Fazit und Ausblick

Ein Benutzer eines Information Retrieval Systems hat nie die Perspektive einer großangelegten Evaluierurungsstudie wie TREC oder CLEF sondern sieht immer nur die Performanz eines Systems für seine aktuelle Fragestellung. Der Robsut Task bei CLEF 2006 stellt die Benutzerperspektive in das Zentrum und evaluiert Systeme danach.

Der Robust Task stellt über die CLEF Initiative 2006 hinaus ein wertvolles Forschungsinstrumentarium zur Verfügung. Die Kollektion kann weiter benutzt werden. Auch die erzielten Ergebnisse werden weiterhin ausgewertet.

Danksagung

Der *Robust Task* wäre nicht möglich gewesen ohne die Hilfe zahlreicher Personen. Im Vorfeld und begleitend hat Ellen Voorhees (NIST, USA) bei allen Entscheidungen beraten. Ein *Robust Task Committee* hat die das Design der Aufgaben betreut. Dazu gehörten: Donna Harman (NIST, USA), Carol Peters (ISTI-CNR, Italien), Jacques Savoy (Universität Neuchâtel) und Gareth Jones (Dublin City Uuniversity). Dank der Hilfe von Giorgio di Nunzio und Nicola Ferro (Universität Padua) konnte der *Robust Task* überhaupt durchgeführt werden. Sie haben beim Task Design mitgewirkt und die CLEF Infrastruktur (das DIRECT System) angepasst. Zuletzt sei allen Teilnehmern gedankt, die Ihre Zeit in die Experimente investiert haben, um so zum Gelingen des *Robust Tasks* beizutragen.

Literaturverzeichnis

Braschler Martin (2003): CLEF 2002 - Overview of Results. In Evaluation of Cross-Language Information Retrieval Systems. Third Workshop of the Cross Language Evaluation Forum 2003, Trondheim. Berlin et al.: Springer (LNCS) Preprint http://www.clef-campaign.org.

Braschler, Martin; Peters Carol (2004): Cross-Language Evaluation Forum: Objectives, Results, Achievements. Information Retrieval. no. 7. 7-31.

Buckley, Chris; Voorhees, Ellen (2005): Retrieval System Evaluation. In: TREC: Experiment and Evaluation in Information Retrieval. Cambridge & London: MIT Press. pp. 53-75.

Buckley, Chris (2004): Why current IR engines fail. In: Proceedings of the 27th annual international conference on research and development in information retrieval (SIGIR 2004), New York: ACM Press. S. 584-585.

Cronen-Townsend, S.; Zhou, Y.; Croft, W. (2002): Predicting Query Performance. Proc Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR '02) (Tampere, Finland, Aug. 11-15, 2002). ACM Press, 299-306.

Eguchi K, Kando Noriko; Kuriyama K (2002): Sensitivity of IR Systems Evaluation to Topic Difficulty. In Araujo CPS and Rodríguez MG (eds.). Proc Third International Conference on Language Resources and Evaluation (LREC) (Las Palmas de Gran Canaria, Spain, May 29-31), 585-589.

Goni-Menoyo, José; Gonzalez-Cristobal, José; Vilena-Román, Julio (2006): Report of the MIRACLE teach for the Ad-hoc track in CLEF 2006. In: Nardi, Alessandro; Peters, Carol; Vicedo, José Luis (Eds.): CLEF 2006 Working Notes.

Harman, Donna; Voorhees, Ellen (1997): Overview of the Sixth Text REtrieval Conference. In Harman D and Voorhees E (ds.). The Sixth Text REtrieval Conference (TREC-6). NIST Special Publication, Gaithersburg, Maryland, 1997, http://trec.nist.gov/pubs/

Harman, Donna; Buckley, Chris (2004): RIA and 'Where can IR go from here?'. In: ACM SIGIR Forum, vol. 38, (2). S. 45-49 .

Kwok, K (2005): An Attempt to Identify Weakest and Strongest Queries. In: SIGIR Workshop Predicting Query Difficulty. 2005. http://www.haifa.il.ibm.com/sigir05-qp

Mandl, Thomas.; Womser-Hacker, Christa (2005): The Effect of Named Entities on Effectiveness in Cross-Language Information Retrieval Evaluation. Proc ACM Symposium on Applied Computing (SAC). Santa Fe, New Mexico, USA. March 13.-17. S. 1059-1064.

Mandl, Thomas (2006): Neue Entwicklungen bei den Evaluierungsinitiativen im Information Retrieval. In: Mandl, Thomas; Womser-Hacker, Christa (Hrsg.): Effektive Information Retrieval Verfahren in der Praxis: Proceedings Vierter Hildesheimer Evaluierungs- und Retrievalworkshop (HIER 2005) Hildesheim, 20.7.2005. Konstanz: Universitätsverlag [Schriften zur Informationswissenschaft 45] S. 117-128.

Martinez-Santiago, Fernando; Montejo-Ráez, Atruro; Garcia-Cumbreras, Miguel; Ureña-Lopez, Alfonso (2006): SINAI at CLEF 2006 Ad hoc Robust Multilingual Track: Query Expansion using the Google search engine. In: Nardi, Alessandro; Peters, Carol; Vicedo, José Luis (Eds.): CLEF 2006 Working Notes.

Mothe, Josiane; Tanguy, L. (2005): Linguistic features to predict query difficulty - a case study on previous TREC campaigns. In: SIGIR Workshop Predicting Query Difficulty. http://www.haifa.il.ibm.com/sigir05-qp

Di Nunzio, Giorgio; Ferro, Nicola; Mandl, Thomas; Peters, Carol (2006): CLEF 2006: Ad Hoc Track Overview. In: Nardi, Alessandro; Peters, Carol; Vicedo, José Luis (Eds.): CLEF 2006 Working Notes.

Peters, Carol Braschler, Martin, Gonzalo, Julio; Kluck, Michael (2003) (eds.): Evaluation of Cross-Language Information Retrieval Systems. Third Workshop of the Cross Language Evaluation Forum 2002, Rome. Berlin et al.: Springer [Lecture Notes in Computer Science 2785]

Peters, Carol; Braschler, Martin, Gonzalo, Julio; Kluck, Michael (2004) (eds.):Evaluation of Cross-Language Information Retrieval Systems. Third Workshop of the Cross Language Evaluation Forum 2003, Trondheim. Berlin et al.: Springer [Lecture Notes in Computer Science] Preprint http://www.clef-campaign.org.

Savoy, Jaques; Abdou, Samir (2006): UniNE at CLEF 2006: Experiments with Monolingual, Bilingual, Domain-Specific and Robust Retrieval. In: Nardi, Alessandro; Peters, Carol; Vicedo, José Luis (Eds.): CLEF 2006 Working Notes.

Tomlinson, Stephen (2006): Comparing the Robustness of Expansion Techniques and Retrieval Measures. In: Nardi, Alessandro; Peters, Carol; Vicedo, José Luis (Eds.): CLEF 2006 Working Notes.

Vilares, Jesús; Oakes, Michael, Tait, John (2006): CoLesIR at CLEF 2006: Rapid Prototyping on an N-gram-based CLIR System. In: Nardi, Alessandro; Peters, Carol; Vicedo, José Luis (Eds.): CLEF 2006 Working Notes.

Voorhees, Ellen (2005a): The TREC robust retrieval track. In: ACM SIGIR Forum 39 (1) 11-20. Voorhees, Ellen (2005a): The TREC robust retrieval track. In: ACM SIGIR Forum 39 (1) 11-20.

Voorhees, Ellen (2005b): Overview of the TREC 2005 Robust Retrieval Track. In: Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005) Gaithersburg, Maryland, November 15-18, 2005. http://trec.nist.gov/pubs/trec14/t14_proceedings.html

Womser-Hacker, Christa (1996): Das MIMOR-Modell. Mehrfachindexierung zur dynamischen Methoden-Objekt-Relationierung im Information Retrieval. Universität Regensburg, Habilitationsschrift.

Zazo, Angel; Figuerola, Carlos, Berrocal, José (2006): REINA at CLEF 2006 Robust Task: Local Query Expansion Using Term Windows for Robust Retrieval. In: Nardi, Alessandro; Peters, Carol; Vicedo, José Luis (Eds.): CLEF 2006 Working Notes.